Exploring equivalence domain in nonlinear inverse problems using Covariance Matrix Adaption Evolution Strategy (CMAES) and random sampling

Author(s):
Grayver, Alexander V.; Kuvshinov, Alexey

Publication Date:
2016-05

Permanent Link:
https://doi.org/10.3929/ethz-b-000114098

Originally published in:

Rights / License:
In Copyright - Non-Commercial Use Permitted
Exploring equivalence domain in non-linear inverse problems using Covariance Matrix Adaption Evolution Strategy (CMAES) and random sampling

Alexander V. Grayver$^1$ and Alexey V. Kuvshinov$^1$

$^1$Institute of Geophysics, ETH Zürich, Sonneggstrasse 5, 8092 Zürich, Switzerland.
E-mail: agrayver@erdw.ethz.ch

8 February 2016

**SUMMARY**

This paper presents a methodology to sample equivalence domain (ED) in non-linear PDE-constrained inverse problems. For this purpose, we first applied state-of-the-art stochastic optimization algorithm called Covariance Matrix Adaptation Evolution Strategy (CMAES) to identify low misfit regions of the model space. These regions were then randomly sampled to create an ensemble of equivalent models and quantify uncertainty. CMAES is aimed at exploring model space globally and is robust on very ill-conditioned problems. We show that the number of iterations required to converge grows at a moderate rate with respect to number of unknowns and the algorithm is embarrassingly parallel. We formulated the problem by using the generalized Gaussian distribution. This enabled us to seamlessly use arbitrary norms for residual and regularization terms. We show that various regularization norms facilitate studying different classes of equivalent solutions. We further show how performance of the standard Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm can be substantially improved by using information CMAES provides. This methodology was tested by using individual and joint inversions of Magnetotelluric, Controlled-source Electromagnetic (EM) and Global EM induction data.
1 INTRODUCTION

Geophysical methods are the primary tool in studying Earth’s interior. They aim at estimating distribution of physical properties in subsurface. This requires solving inverse problems (Aster et al. 2011). Inverse problems rarely have the unique solution and it is more appropriate to consider a set of solutions which are physically plausible and explain data within given error bounds. This set represents the equivalence domain (ED) and serves as a measure of inverse solution uncertainty. Methods aimed to explore ED are attracting widespread interest in geophysics. In this paper, we propose a methodology to probe ED using a state-of-the-art evolution strategy algorithm and random or Markov chain Monte Carlo (MCMC) sampling algorithms.

In what follows, we concentrate on non-linear PDE-constrained inverse problems. They can be tackled by means of deterministic or probabilistic methods. Within deterministic approach, non-linear inversion is posed in form of a minimization problem and commonly solved by using derivative-based methods such as quasi-Newton or Gauss-Newton. The prevalence of derivative-based methods is twofold: (i) for PDE-constrained problems, calculation of derivatives is efficiently done by means of the adjoint method (e.g. Pankratov & Kuvshinov 2015b); (ii) convergence rates of these methods show little dependence on a number of unknowns provided that the minimized functional fulfills differentiability conditions (Nocedal & Wright 1999). These factors make derivative-based methods computationally efficient and thus highly suitable for problems with large number of unknowns. As a downside, they can only find local minima, and highly depend on an initial guess. Within this approach, a model can be appraised using linearized analysis (Menke 2012) which is valid only in the vicinity of the found solution. Despite this, such analysis can still be useful and insightful (Grayver et al. 2014; Fichtner & Leeuwen van 2015). Another difficulty arises when non-smooth problems are to be solved. For instance, usage of $L_1$-norm generally destroys differentiability. Although efficient methods to mitigate this limitation ex-
ist (Farquharson & Oldenburg 1998; Borsic & Adler 2012), they require additional computational resources and can affect convergence rate.

In probabilistic paradigm, one can solve an inverse problem using Bayesian inference or stochastic optimization methods. In contrast to derivative-based methods, they can be considered as global methods and applied to non-smooth and non-convex functions. Bayesian inference describes the solution in form of a posteriori probability density function (PDF) (Tarantola 2005). The question of uncertainty quantification (UQ) can then be addressed by calculating statistical moments or marginal probabilities for a posteriori PDF. The Markov chain Monte Carlo (MCMC) methods are typically used to reconstruct a posteriori PDF. However, they suffer from the “curse of dimensionality” (Curtis & Lomax 2001) and pose formidable challenges as a number of unknowns grows. Much research in recent years has focused on developing approaches aimed at reducing dimensionality of a problem or making sampling process more efficient. Among such techniques are adaptive and ad-hoc model parameterizations (Sambridge & Mosegaard 2002; Laloy et al. 2012; Lochbühler et al. 2014), PCA-based sampling (Malinverno 2002; Dosso & Wilmut 2008), multi-chain MCMC (Laloy & Vrugt 2012; Sambridge 2013; Ray et al. 2013) and the use of local derivative information (cf. Martin et al. 2012; Bui-Thanh & Girolami 2014; Pankratov & Kuvshinov 2015a). Despite the substantial progress, application of MCMC for non-linear problems with large number of unknowns remains very challenging.

Alternatively, inverse problems can be solved by using stochastic optimization methods (Sen & Stoffa 2013). In this work, we applied a stochastic optimization technique called Evolution Strategy with Covariance Matrix Adaptation (CMAES). CMAES was shown to work well on a wide class of functions (Hansen et al. 2010; Rios & Sahinidis 2013). It exhibits remarkable robustness on ill-conditioned and noisy problems and outperforms other popular stochastic methods such as Genetic Algorithms and Particle Swarm Optimization in functional optimization (Auger et al. 2009; Hansen et al. 2011) and PDE-constrained parameter estimation (Arsenault et al. 2013; Elshall et al. 2015) problems. Such a robustness is favourable since geophysical inverse problems are usually ill-conditioned. Additionally, CMAES can incorporate linear constraints (Diouane et al. 2015).
Despite advantages mentioned above, the use of CMAES in geophysics is not common, although it recently generated some interest among researchers in geophysics (Alvers et al. 2013; Fonseca et al. 2014; Diouane 2014; Shen et al. 2015). These studies, however, focused on using CMAES as a global minimization method only. However, we are additionally interested in exploring ED and quantifying solution uncertainty. Even though stochastic methods exhibit many similarities with MCMC (Tarantola 2005; Mueller 2010), their primary goal is to minimize a given functional and hence during sampling of the model space they lean toward regions of low values. In contrast to MCMC, they do not sample from the stationary distribution and estimators based on these methods tend to be biased (Sambridge & Mosegaard 2002; Cerv et al. 2007). Therefore, we exploit advantages of CMAES and combine it with random sampling or MCMC to explore ED. As an application example, we used 1D inversion of electromagnetic data. Nevertheless, the presented methodology and code are versatile, and can be applied to essentially any PDE-constrained inverse problem.

We note that this study does not address the questions of model selection and regularization parameter choice. The former question can be addressed by using the trans-dimensional MCMC (Malinverno 2002; Sambridge et al. 2006; Minsley 2011; Ray et al. 2014) or heuristically by Bayesian information criterion (BIC) (Guo et al. 2011). There are many methods aimed at helping to choose the regularization parameter value within deterministic approach (Farquharson & Oldenburg 2004). If standard deviations are well known, the approach adopted by Occam’s inversion (Constable et al. 1987) can be used, where local univariate search of the regularization parameter is performed using discrepancy principle. Alternatively, hierarchical Bayesian approach can be used to co-estimate probability function of the regularization parameter (Rosas-Carbajal et al. 2013). Note that BIC and discrepancy principle could be used with stochastic methods, but would hardly influence conclusions drawn in this work, hence we defer their investigation to future works.

2 THEORY AND METHOD

Following Bayesian probabilistic formulation, let us define the posterior PDF of a model \( m \in \mathbb{R}^M \) given measured data \( d \in \mathbb{R}^N \) by
\begin{equation}
P_{\text{post}}(m|d) \propto P_{\text{like}}(d|m)P_{\text{prior}}(m),
\end{equation}

with \( N, M \) being number of measurements and unknown model parameters, respectively. Adopting generalized Gaussian distribution (GGD) model (Tarantola 2005), likelihood and prior probabilities in equation (1) can be written as

\begin{equation}
P_{\text{like}}(d|m) \propto \exp \left( -\frac{1}{p_d} \sum_{i=1}^{N} |w_i (f(m) - d)|^{p_d} \right)
\end{equation}

and

\begin{equation}
P_{\text{prior}}(m) \propto \exp \left( -\frac{1}{p_m} \sum_{i=1}^{M} |l_i (m - m_{\text{ref}})|^{p_m} \right),
\end{equation}

where \( f(m) : \mathbb{R}^M \rightarrow \mathbb{R}^N \) is a forward modeling operator, \( w_i \) and \( l_i \) denote \( i \)th rows of the data \( W \) and model \( L \) weighting matrices, respectively (Verdoolaege & Scheunders 2012); \( m_{\text{ref}} \) is a reference model. In this paper, \( W \) is a diagonal matrix with \( W_{ii} = \frac{1}{(\delta d_{pd})_i} \), where \( \delta d_{pd} \) indicates generalized standard deviation (Borsic & Adler 2012). The orders of GGD for likelihood and prior terms are given by \( p_d \) and \( p_m \), respectively.

The maximum posterior probability model, \( m_{\text{MAP}} \), can be obtained by solving the following optimization problem

\begin{equation}
m_{\text{MAP}} = \max_m P_{\text{post}}(m|d) = \arg\min_m \phi(m),
\end{equation}

with the cost function given by

\begin{equation}
\phi(m) = \frac{1}{p_d} \sum_{i=1}^{N} |w_i (f(m) - d)|^{p_d} + \frac{\beta}{p_m} \sum_{i=1}^{M} |l_i (m - m_{\text{ref}})|^{p_m},
\end{equation}

where regularization matrix \( L \) can be discrete form of first derivative operator or identity matrix and \( \beta \) is the regularization parameter.

### 2.1 CMAES

In this work, we used CMAES to solve problem (4). This section discusses key features of the CMAES, whereas the full algorithm is given in the Appendix A. A generic evolution strategy is shown in Algorithm 1. At iteration \( g \) (also called generation), a population consisting of \( \lambda \) models is sampled from a distribution \( P(m|\theta) \), where \( \theta \) are distribution parameters such as mean
vector and covariance matrix. After evaluating all sampled models, the distribution parameters are updated in a function $F(\cdot)$, and algorithm proceeds. It is the choice of distribution and the update function $F(\cdot)$ which determine the gist of CMAES. CMAES adopts multivariate normal distribution model, hence

$$m_i^g \sim \mathcal{N}(\bar{m}^g, C^g), \quad i = 1, \ldots, \lambda.$$  

\[ (6) \]

\textbf{Algorithm 1} An evolution strategy template to minimize $\phi(m)$

1: Initialize distribution parameters $\theta^0$

2: Set iteration number $g = 0$

3: \textbf{repeat}

4: Draw $\lambda$ sample models from $P(m|\theta^g) \rightarrow m_1, \ldots, m_\lambda$

5: Evaluate $\phi(m_i) \quad i = 1, \ldots, \lambda$

6: Update parameters $\theta^{g+1} \leftarrow F(\theta^g, m_1, \ldots, m_\lambda, \phi(m_1), \ldots, \phi(m_\lambda))$

7: $g \leftarrow g + 1$

8: \textbf{until} termination criterion is met

More specifically, we used $(\mu/\mu_w, \lambda)$–CMAES algorithm. This notation indicates that the algorithm selects $\mu$ best models out of $\lambda$ generated candidates at iteration $g$ and their weighted combination is used to update distribution parameters. The way normal distribution parameters are updated is shown in the Appendix A

From a user perspective, only two parameters need to be predefined. Namely, initial guess $m_0$ and step size $s$ (cf. Algorithm (4) in Appendix A). Larger $s$ values let algorithm to sample model space more widely, although may require more iterations to find the minimum. Too small values will limit algorithm to sample in the vicinity of the initial guess and therefore hide the global minimum. A reasonable choice for the initial step size is $s = (b - a)/3$ (Hansen & Ostermeier 2001), where $a, b$ specify bounds of the parameter values, within which the solution is expected to lie. In principle, the algorithm controls step size adaptively and therefore in practice one rarely needs to fine-tune the step size. Additionally, a user may specify population size $\lambda$ which default
value is chosen as
\[ \lambda = 4 + \lfloor 3 \log M \rfloor, \tag{7} \]
where \( \lfloor \cdot \rfloor \) is the floor function. Equation (7) is a reasonable choice for initial trials (Hansen & Ostermeier 2001). Unless otherwise stated, equation (7) was used to choose population size \( \lambda \) in all experiments and examples.

Numerous modifications of original Algorithm (4) exist. In this work, we applied two of them. First, active CMAES (Hansen & Ros 2010) which further improves performance of CMAES on ill-conditioned problems by using more elaborated method to update covariance matrix. Both classic and active CMAES use full covariance matrix internally. To draw samples from multivariate normal distribution in Algorithm (4) and to rotate weighted sum \( \langle m \rangle_w \) to principal axes (step 12 in Algorithm 4), one needs to calculate eigendecomposition of the covariance matrix
\[
C_{CMA} = V \Lambda V^T. \tag{8}
\]
This results in \( \mathcal{O}(M^3) \) time complexity. In practice, it is possible to reduce this to \( \mathcal{O}(M^2) \) (Hansen 2006). Nevertheless, this still limits feasibility of CMAES for large \( M \).

Therefore, we tested another modification called VDCMA (Akimoto et al. 2014). It replaces full covariance matrix \( C_{CMA} \) with
\[
C_{CMA}^{VD} = D(I + vv^T)D^T, \tag{9}
\]
where \( D \) is a diagonal scaling matrix and \( v \) a principal rotation vector. This representation leads to \( \mathcal{O}(M) \) time and memory complexity of the algorithm. Despite reduced amount of information in covariance matrix, VDCMA was shown to work well for problems with up to \( 10^4 \) unknowns (Akimoto et al. 2014).

As an example, Figure 1 illustrates evolution paths, that is mean models at every iteration, taken by the active CMAES and models produced by Gauss-Newton method for a non-linear least-squares \( (p_d = 2) \) fitting problem without regularization (Menke 2012)
\[
f(m) = \sin(20m_1x) + m_1m_2, \quad 0 \leq x \leq 1. \tag{10}
\]
For this experiment, true data were contaminated with normal noise \( \sim \mathcal{N}(0, 0.4) \) and colour panels
Figure 1. Upper plots: true (black solid curve) and noise-contaminated (black circles) data for problem (10). Responses calculated for final solutions are given in color. Note that for CMAES all response curves look identical since it converged to the same solution independent of an initial guess. Lower plots: optimization paths of Gauss-Newton (a) and active CMAES (b) algorithms for three different initial guesses: [1.01, 0.9] (yellow + circles), [0.98, 0.9] (red + squares), [0.1, 0.4] (magenta + triangles). Root mean square error is shown in colour. White triangle depicts position of the true solution.

show RMS error calculated using noisy data. These noisy data were then used to estimate parameters \([m_1, m_2]\) by means of CMAES and Gauss-Newton method. As can be seen, even initial points which are close to each other may hinder convergence of the Gauss-Newton method to the global minimum, whereas CMAES recovers global minimum independent of an initial guess in less than 40 iterations.
2.2 Norms of the residual and regularization terms

In this section, we discuss the choice of $p_d, p_m$ in equations (2,3) and implications for the solution of equation (4). As can be seen from these equations, for orders $p_d = p_m = 2$ normal multivariate distribution is recovered. Therefore, operators $W^T W$ and $L^T L$ correspond to the inverse of data and model covariance matrices, respectively. For other orders, these matrices do not represent covariance matrices anymore and are conventionally called scatter matrices (Pascal et al. 2013). Care is required in choosing data uncertainty values in $W$ for values $p_d \neq 2$ since the normal standard deviation $\delta d_2$ is related to other orders as $\delta d_{p_d} = \delta d_2 \sqrt{\frac{\Gamma(1/p_d)}{\Gamma(3/p_d)}}$ (Sharifi & Leon-Garcia 1995). Therefore, one has to convert normal data uncertainties accordingly.

When searching the MAP solution, $p_d$ and $p_m$ essentially determine norms of the residual and regularization terms in equation [5]. From this perspective, the choice of $p_d = p_m = 2$ corresponds to the conventional non-linear least-squares problem with Tikhonov-type regularization. The mea-
sure of misfit in this case is the well-known root mean square error (RMSE)

$$\delta_2(m) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{(f(m)_i - d_i)^2}{(\delta d_i)_i^2}}, \quad (11)$$

which ≈ 1 if responses fit data within uncertainties and residuals are normally distributed and have zero mean. Stacked Fourier transformed EM data will typically have Gaussian residuals and the choice of $p_d = 2$ is justified in the majority of cases (Wheelock et al. 2015). However, if this does not hold, the use of other distributions (and norms) may be more appropriate. In particular, for $p_d = 1$, Laplace distribution for residuals is obtained where

$$\delta_1(m) = \frac{1}{N} \sum_{i=1}^{N} \frac{|f(m)_i - d_i|}{(\delta d_i)_i}, \quad (12)$$
is the appropriate measure of data misfit, also called mean absolute error (MAE). Assuming that residuals obey Laplace distribution makes inversion robust against outliers (due to longer tails, see Figure 2) and MAE is a more reliable measure (Willmott & Matsuura 2005). This property was found to be useful in geophysics (Farquharson & Oldenburg 1998; Walker & Jackson 2000). Other orders and corresponding norms are permitted within our formulation, but are rarely (if at all) used in practice. One approach that may prove useful is switching norms during the process. This can be justified since initial residuals can be far from Gaussian because of an initial model being far from the MAP solution. In this case, using more permissive distribution at initial stages and switching to Gaussian distribution later can improve convergence and reduce the risk of model artifacts.

For regularization term, $p_m = 2$ corresponds to the widely used $L_2$ regularization. If $L$ in equation (5) approximates a first first-order derivative operator, we get smoothing regularization. Accordingly, $p_m = 1$ corresponds to the $L_1$ regularization favoring compact models, possibly with sharp discontinuities in a model. In addition to these widely used choices, we have also experimented with other values of $p_m$ in this work.

Recognizing that distribution of likelihood and prior PDFs are tightly connected with norms in equation (5), the choice of $p_d$ and $p_m$ allows us to unravel different classes of feasible solutions.
2.3 Exploring equivalence domain

Solution of equation (11) is a single model. However, given noise in data and a finite number of measurements, it is more appropriate to consider a set of admissible solutions, \( S_d \), which explain data within uncertainties, that is

\[
S_d = \left\{ m : \hat{\delta}_{p_d}(m) \leq \hat{\delta}_{p_d} \right\},
\]

where \( \hat{\delta}_{p_d} \) is a predefined data misfit threshold.

For linear problems, \( S_d \) is a simply connected set (Fernández-Martínez et al. 2014b), that is it consists of a single low misfit region which has elliptical shape, although this hyper-ellipsoid degenerates and becomes unbounded for underdetermined or rank-deficient problems. In case of non-linear problems, \( S_d \) is not simply connected (there may be several disconnected low misfit regions) and may or may not be bounded (Fernández-Martínez et al. 2014a). Many models in ED will be physically implausible and one generally wants to impose constraints and restrict this set to a smaller class of physically meaningful solutions.

Probabilistic approach allows to do that by using a non-flat prior PDF (Rosas-Carbajal et al. 2013). Within deterministic approach, this is usually done by introducing a regularization. This requires \( m_{MAP} \) to be from a correctness set \( S_m \) induced by a prior PDF (for probabilistic inversion) or stabilizing functional (for deterministic inversion).

Without loss of generality, we define ED to be the following set

\[
S_{ED} = S_d \cap S_m.
\]

Note that in general a correctness set \( S_m \) depends on the choice of \( p_m \), regularization operator \( L \), reference model \( m_{ref} \) and regularization parameter \( \beta \). A proper choice of these parameters ensures that \( S_{ED} \neq \emptyset \). In other words, that there is at least one model fitting the data for the given regularization settings.

As an output, CMAES returns \( m_{MAP} \) as well as a covariance matrix estimate, \( C_{CMA} \). This estimate approximates the true covariance matrix up to a scalar and some random perturbations (due to stochastic nature of the algorithm). Therefore, one can readily appraise uncertainty locally. However, exploring ED using essentially elliptical approximation in the vicinity of \( m_{MAP} \)
is not sufficient since in general ED is not a simply connected set and may have complex shape (Fernández-Martínez et al., 2013). To perform a more comprehensive exploration of ED and retain practical feasibility, we devised a two steps methodology.

Denote by \( S_{\text{low}} \) the following set of models

\[
S_{\text{low}} = S_{\text{CMAES}} \cap S_d,
\]

where \( S_{\text{CMAES}} \) is a set of all candidates generated by CMAES during the evolution. Although CMAES tries to converge to the \( m_{\text{MAP}} \) model independent of an initial guess, the optimization path taken depends on an initial guess and a random number generator (RNG) seed (Park & Miller, 1988). By changing one or both of them, CMAES will probe different parts of the ED during evolution. Taking this into account, the first step is to construct a \( S_{\text{low}} \) by running the Algorithm (2).

**Algorithm 2 Constructing \( S_{\text{low}} \)**

1. Choose \( p_d, p_m \), regularization parameter \( \beta \), misfit threshold \( \hat{\delta}_{p_d} \), bounds of parameters \( m_{\text{min}}, m_{\text{max}} \) and number of restarts \( n_r \).
2. Set \( S_{\text{low}} = \emptyset \).
3. for \( i \leftarrow 1, n_r \) do
4. Draw random starting model \( m_0 \sim U(m_{\text{min}}, m_{\text{max}}) \) and RNG seed
5. Solve problem (4) using CMAES with these settings
6. \( S_{\text{low}}^i = S_{\text{CMAES}}^i \cap S_d \)
7. \( S_{\text{low}} \leftarrow S_{\text{low}} \cup S_{\text{low}}^i \)

Set \( S_{\text{low}} \) contains low misfit solutions sampled during evolution process from \( n_r \) restarts with randomly chosen initial models and random generator seeds. It typically consists of many similar solutions which tend to cluster around \( m_{\text{MAP}} \) since this is the point which algorithm is supposed to find. For large \( n_r \), this set can become representative of ED, but this quickly becomes rather time consuming, especially for large \( M \) or high dimensional PDEs. On the other hand, to get an idea about solution uncertainty, we would like to explore the most different equivalent solutions. To this end, a small number \( k \) of low misfit candidates which are approximately evenly distributed
within $S_{low}$ with respect to $m_{MAP}$ were selected. When selecting these candidates, it is important to account for rotation and shape of the misfit function around $m_{MAP}$. Covariance matrix contains exactly this type of information. Therefore, we adopted Mahalanobis distance for this task defined as

$$d(m) = \sqrt{(m - m_{MAP}) C_{MAP}^{-1} (m - m_{MAP})},$$ \hspace{1cm} (16)$$

Further, we randomly sample ED around models from $S_{low}$ and $m_{MAP}$. To increase efficiency of the sampling, we also employed $C_{MAP}$, i.e. covariance matrix at maximum a posteriori. This approach has often been exploited in Bayesian inference [Kaipio et al. 2000; Malinverno 2002; Dosso & Wilmot 2008] since topography of the model space may become very elongated along some directions for ill-conditioned or/and numerically rank-deficient problems. Therefore, performing sampling in the space rotated to the principal axes of $C_{MAP}$ is more efficient.

However, $C_{CMA}$ cannot be readily used as an approximation of $C_{MAP}$. Being an optimization algorithm, CMAES tends to shrink covariance matrix [Shir et al. 2014], although the shape and rotation of the covariance hyper-ellipsoid are generally well recovered. Therefore, to be able to use it for random sampling, $C_{CMA}$ matrix needs to be rescaled. To do this, first recall that at $m_{MAP}$ the covariance matrix is (asymptotically) given by the mathematically well-posed inverse of the cost functional Hessian matrix. We therefore choose a scaling factor so that the quadratic approximation is recovered. To this end, we perturbed $m_{MAP}$ along one principal axis as follows

$$m^* = m_{MAP} + v \delta p,$$ \hspace{1cm} (17)$$

where $v$ is the dominant eigenvector of the covariance matrix $C_{CMA}$ and $\delta p$ is a small number. The rescaled covariance matrix is then obtained by virtue of Taylor expansion as

$$C_{MAP} \approx \frac{(m^* - m_{MAP})^T C_{CMA}^{-1} (m^* - m_{MAP})}{2 (\phi(m^*) - \phi(m_{MAP}))} C_{CMA}.$$ \hspace{1cm} (18)$$

As can be seen, this rescaling is done at the cost of one additional forward calculation.

Note that for non-smooth problems, such as $p_d = 1, p_m = 1$, the use of covariance matrix for random sampling becomes somewhat contradictory. Indeed, strictly speaking, no covariance matrix for non-Gaussian case can be derived. Therefore, when $p_m, p_d = 1$ we used only main
diagonal of the covariance matrix, that is $C_{CMA} = \text{diag}(C_{CMA})$ and discarded linear interrelations given by non-diagonal terms.

**Algorithm 3** Constructing a subset of $S_{ED}$, $S_{d,m}$

1: Construct $S_{low}$ using Algorithm (2).

2: Select a subset $S_{mean} \in S_{low}$ of $k$ models which are evenly distributed around $m_{MAP}$ and add $m_{MAP}$ itself to $S_{mean}$.

3: Obtain $C_{MAP}$ using equation (18).

4: Set $S_{d,m} = \emptyset$

5: repeat

6: Choose randomly $\bar{m} \in S_{mean}$

7: Sample random model $m \sim \mathcal{N}(\bar{m}, C_{MAP})$

8: if $\delta_{pd}(m) \leq \hat{\delta}_{pd}$ then

9: Add $m$ to $S_{d,m}$

10: until termination criterion is met

With this at hand, we can finally construct a subset of $S_{ED}$ following procedure summarized in Algorithm (3). Termination criteria in the algorithm are one of the following:

- Maximum allowed number of forward solutions is reached;
- Required number of accepted samples, $n_A$, is reached.

After termination, we obtain $S_{d,m}$ – a set of models which fit data within given uncertainties and are physically plausible (under the constraints prescribed by the regularization and optional parameter bounds). This subset can be used to quantify uncertainty.

The example shown in Figure 3 illustrates random sampling of ED. For this experiment, ED is defined as a set of all models with $\delta_2(m) \leq 2.5$. This threshold was chosen with an intention to make ED a not simply connected set. Five restarts were used in Algorithm (2). Due to choosing multiple low-misfit points (gray squares in Figure 3), we are able to sample disconnected regions of the ED uniformly. This is especially important when dealing with multi-modal posterior distributions.
Figure 3. Sampling ED for problem (10) using Algorithms (2) and (3). (a) Shown are boundaries of ED (yellow solid contours), \( m_{MAP} \) (white triangle) and scaled covariance ellipse (dashed magenta line) returned by CMAES. A subset of \( k = 50 \) low misfit models which are uniformly distant from the MAP point are shown with gray squares. (b) \( n_A = 1000 \) equivalent models randomly sampled from the ED drawn using Algorithm (3).

In general, the coverage of ED will largely depend on a number of restarts, \( n_r \), in Algorithm (2) and a number of accepted samples, \( n_A \), in Algorithm (3). These parameters remain problem dependent.

In summary, the presented approach of ED estimation consists of two steps:

- Find MAP model and identify low-misfit regions of the parameter space using Algorithm (2). The key ingredient of this step is CMAES. As a result, we obtain the MAP model, covariance matrix and a set of low-misfit solutions probed during optimization.
- Using information obtained at the previous step, draw samples from the ED by means of random sampling (Algorithm (3)).

3 PARALLEL IMPLEMENTATION

One advantage of the Algorithms (2-4) is their parallel nature. Often the most time-consuming step 9 of CMAES Algorithm (4) can be done in parallel with minimal communication involved, since forward problems for all candidates in a population can be solved independently. The only
data to be communicated between MPI processes are generated candidates and fit function values once calculations are done. Considering that the recommended population size in CMAES grows rather moderately with $M$ of a problem (cf. equation 7), a tractable number of CPUs is sufficient to reduce the cost of a single iteration down to a cost of a single forward problem plus cost of mean vector and covariance matrix updates. The same holds for Algorithm 2 where multiple chains can be run in parallel and produced sets can be combined afterward. Finally, random sampling in the Algorithm 3 is also easy to parallelize with multiple independent chains. These considerations make it a favourable algorithm to be run on distributed platforms widely available nowadays. In this case, significant computational savings will be obtained for applications where forward operators are expensive.

We used CMAES implementation from the open-source libcmaes library (Benazera & Hansen 2015) and created a custom strategy to enable parallelism as discussed in the previous paragraph.

4 NUMERICAL EXPERIMENTS

In this section, we demonstrate various aspects of the presented algorithms. We invert data for a variety of geo-electromagnetic methods which are used on very different scales: from global studies to local exploration surveys. By doing this, we want to highlight not only challenges they share, but also efficiency of the proposed algorithms for a wide range of problems. Unless otherwise stated, we used active CMAES with default parameters for all experiments.

4.1 Non-linearity: a motivating example

We begin by considering a simple problem of identifying thickness and conductivity of a thin resistive layer embedded in a conductive background using controlled-source electromagnetic method (CSEM) (Key 2009). However simple, it will be shown that this example already poses challenges which are typical for more realistic problems.

The true model for this experiment is shown in Figure 4. Two unknown parameters are logarithms of thickness and conductivity of the thin resistive layer. The top of the layer was set to the true value and remained fixed. Amplitudes and phases of $E_y, B_x$ components at the frequencies
Exploration of equivalences in inverse problems

Figure 4. True model used in the synthetic test (not to scale). Positions of receiver and electric dipole transmitters are denoted by the triangle and stars, respectively.

of 1 and 0.1 Hz were calculated for a single receiver at the sea floor and 22 transmitters placed 500 – 11000 m away from the receiver with the spacing of 500 m. In total, 176 real-valued values contaminated with 7% Gaussian noise and 7% error floor were inverted. No regularization was used.

Figure 5(a) shows misfit function landscape for this minimization problem. As anticipated

Figure 5. (a) Evolution paths of the active CMAES plotted over misfit function landscape for reservoir problem with two unknowns. (b) RMSE values along the dashed line shown in (a). (c) RMSE versus iteration number for the evolution path shown in (a).
from theory (Fernández-Martínez et al. 2014a), the low misfit regions resemble meandering valley shape with (in this case) two disconnected local minima. Additionally, what makes this problem challenging for derivative-based methods is the existence of a ridge followed by the flat plateau toward higher conductivity values (see Figure 5b). Starting on the “wrong” side of the ridge will preclude derivative based methods from converging.

In contrast, CMAES converges quickly in this case (Figure 5b). For this run, \( m_0 = (0.1, 10) \) was used as a starting point and it took 45 iterations to reach RMSE of one (Figure 5c) (RMSE = 1.26 was reached in 30 iterations). Figure 6(a) shows equivalence domain (defined for all models with \( \hat{\delta}_2 \leq 1.5 \)). It consists of a single slightly bended region. Note that MAP solution is located away from the center of this domain indicating non-symmetric nature of the posterior PDF even on the small scale. To explore ED, we ran random sampling following methodology presented in Section 2.3. Ten restarts were performed with \( k = 50 \) and \( n_A = 4000 \) equivalent solutions shown in Figure 6(a) with yellow dots were generated. As can be seen, sampled solutions cover ED rather uniformly. Figure 6(b) plots MAP and equivalent solutions giving a notion of uncertainty in inverse problem. It is clear that two parameters are also highly correlated. The MAP solution is located at \((0.0097, 96.82)\) and two most distant (in the parameter space) solutions are \((0.006, 60)\) and \((0.013, 141)\). Due to the noise in the data, minimum of the cost functional does not coincide with the true solution \((0.01, 100)\) and, hence, MAP solution is shifted. As was discussed in (Fernández-Martínez et al. 2014a), this will always be the case for noisy problems, highlighting the importance of uncertainty quantification in practice.

### 4.2 Choice of norms

In this section, we study the effect of norms on the minimized functional (5) using global geomagnetic sounding based on Dst source and 1D \( C \)-responses concept (Kuvshinov & Olsen 2006). Recall that 1D \( C \)-response is a frequency dependent complex-valued transfer function which reads

\[
Z = i \omega \mu_0 C,
\]

(19)
where $\omega$ is angular frequency, $\mu_0$ magnetic permittivity of the free space, $Z$ is the 1D impedance. The complex-valued $C$-response has physical dimension of length and its real part provides an estimation of the depth to which the EM field penetrates (Weidelt 1973).

Independent of $p_d$, class of admissible models is determined by a model parameterization. For instance, by choosing a polynomial basis to parametrize a model, one confines all solutions to be from a set of polynomial functions of the given degree. Other choices such as wavelets and periodic functions are possible (Fernández-Muñiz et al. 2015; Hawkins & Sambridge 2015). In this work, we chose a basis of piecewise constant functions – arguably the most widely used model parameterization in geophysics nowadays. In this work, we use a fixed model parameterization and chose a number of layers a priori. One way is to take as few layers as possible and try to fit data by varying both physical parameters and layer thicknesses. The other choice is to deliberately overparameterize a model with many layers of fixed thickness and seek for their physical properties. The latter choice serves as some sort of enrichment of a set of models which can be approximated sufficiently well. However, new degrees of freedom significantly increase ambiguity and to mitigate this, we add regularization. In this experiment, $L$ in equation (5) is the discrete

**Figure 6.** (a) Zoomed in misfit function landscape from Figure 5(a). Shown are equivalence domain (white solid line), MAP solution (magenta triangle), equivalent solutions sampled using Algorithm (5) (yellow points). (b) MAP (dashed line) and equivalent models (gray solid lines).
form of the 1st order derivative operator (Hansen 1998). Additionally, the value of \( p_m \) determines class of plausible models: while \( p_m = 2 \) “prefers” smooth solutions, \( p_m = 1 \) results in compact, but rough models. The real distribution of properties may neither be solely smooth nor rough. This prompts using other “trade-off” metrics, for instance \( p_m = 1.5 \). The aim of this section is to test various norms on two extreme scenarios.

Two models with sharp discontinuities and smooth conductivity variations were generated (black dashed lines in Figures 7a,b). Global \( C \)-responses were calculated for these models at 40 periods in the range between 36 hours and one year. These data were corrupted with 5% Gaussian noise (points with error bars in Figures 7c,d). For this experiment, \( p_d = 2 \). The Earth was parameterized with 71 layers going from the surface to the outer core. To account for loss of resolution with depth, layer thicknesses grow in size from 3 km at the surface to 185.3 km at the core-mantle boundary. Note that we deliberately chose a parameterization such that layer boundaries do not coincide with discontinuities in the true models since in reality one rarely has precise information about layer boundaries. Inversion was performed for both models using a series of \( p_m \) values, namely: 1, 1.5, 2. All other parameters between different runs were identical. Starting model was a homogeneous halfspace model with conductivity of 0.1 S/m.

For the model with sharp conductivity contrasts (Figure 7a), \( L_1 \) norm inversion performed best. It resolved near-surface conductivity and discontinuity at 660 km. The \( L_2 \) norm functional recovered a smooth model. As one can anticipate, using \( L_{1.5} \) norm gives a result in between, without a tendency to create wiggly artifacts typical for \( L_2 \) norm. Application of \( L_1 \) regularization norm to invert data for the smooth conductivity model (Figure 7b) illustrates unsatisfactory performance. While general trend is somehow resolved, the model became oversimplified. Considering that \( L_1 \) aims at finding the most compact model with as few features as possible, this is not surprising. It is deficiency of an assumption about compactness of the sought model which plays the biggest role here. \( L_{1.5} \) and \( L_2 \) norms produce very similar results resolving model well throughout the Earth. If one were to choose a norm which performs best between two of these models, the choice would likely be the \( L_{1.5} \) norm.

Although this study was quite insightful, for all norms used in this experiment, we were able
Figure 7. Black dashed lines illustrate true sharp (a) and smooth (b) synthetic models. 1D responses for sharp (c) and smooth (d) models were calculated at 40 periods and contaminated with 5% Gaussian noise. Inversion results using different $L_{p_m}$ norms are shown with colored solid lines. Legend lists $p_m$ values with final data misfit errors ($\delta_{p_m}$). Since all models fit data equally well, calculated responses are very similar. Therefore, green lines in figures (c) and (d) show predicted data for $p_m = 2$ only. Starting models are depicted in (a, b) with gray dash-dot lines.

to find models which fit data equally well. However, they sometimes look different and could be interpreted differently. This reveals extensive equivalences as well as reflects highly ill-conditioned
nature of the problem, and creates an incentive for methods which enable a proper exploration of these equivalences.

4.3 Convergence study

The comparison of different stochastic optimization methods typically involves plotting number of function evaluations versus function value (Hansen et al. 2010). However, as was discussed in Section 3, CMAES is inherently parallel and with \( \lambda \) CPUs available, the time per iteration boils down to the internal algorithm time plus time required for a single forward problem. The total runtime is therefore proportional to the number of iterations, \( n_g \), required to converge. We therefore prefer to use \( n_g \) as an indicator of algorithm performance (multiplying it by \( \lambda \) yields overall budget of required forward solutions). As with many stochastic optimization methods, it is difficult to give a good \textit{a priori} estimate of \( n_g \). In general, it depends on a function being minimized, algorithm parameters, initial guess and number of unknowns. Considering all these factors goes beyond current scope, we therefore designed an experiment that only studies how the number of unknowns influences convergence for \( p_d = p_m = 1 \) and \( p_d = p_m = 2 \). All other CMAES settings were chosen by default. To alleviate the influence of an initial RNG seed, we ran algorithm 40 times with randomly chosen seeds and calculated mean value with standard deviation.

We used two variants of optimization algorithm discussed in Section 2.1 namely: active modification referred to as aCAMES, as well as VDCMA algorithm which has linear internal time and memory complexity. As an example problem, the model shown in Figure 7(a) was chosen. We parametrized the model with varying number of layers resulting in problems with up to 281 unknowns. The number of iterations till convergence was counted. In this experiment, the algorithm is said to have converged as soon as \( \delta_{p_d}(\bar{m}) \leq 1 \). In other words, if the mean model at the current iteration produces data misfit error \( \leq 1 \). In practice, it is recommended not to stop right upon reaching low misfit solution, but run further to let algorithm do some fine-tuning. This, however, is difficult to quantify rigorously.

As shown in Figure 8(a), for \( L_2 \) norm the number of iterations scales as \( O(M^{1.2}) \). While this
Exploration of equivalences in inverse problems

Figure 8. (a) Number of iterations (generations) versus number of unknowns for $L_2$ norm problem ($p_d = p_m = 2$) using two variants of CMAES and population size $\lambda$. (b) Same as (a), but for $L_1$ ($p_d = p_m = 1$).

growth is higher than linear, it is polynomial and significantly below than quadratic. Moreover, VDCMA which uses reduced covariance matrix and has linear complexity, performed on a par with bona fide CMAES equipped with the full covariance matrix. This encouraging result suggests that the stochastic algorithm can scale to higher dimensional problems which are of great relevance nowadays. Recently, VDCMA was shown to work well on problems with as many as $10^4$ unknowns (Akimoto et al. 2014).

According to the Figure 8(b), CMAES performs slightly worse on $L_1$-norm functional exhibiting $O(M^{1.5})$ growth of number of iterations with respect to the number of unknowns. This is anticipated since it internally exploits multivariate normal distribution model and will naturally perform better on $L_2$ misfit functionals.

One way to reduce number of iterations is to increase population size $\lambda$ in Algorithm (4). Note that due to parallelism over population (Section 3), its size does not affect runtime as long as $\lambda$ CPUs are available. We used constant $\lambda = 100$ instead of equation (7) and results for this test are shown in Figure 8 with squares and rombs. The number of iterations has been systematically reduced at least two times for $L_2$ minimization and even more for $L_1$ norm.
4.4 Exploring equivalence

In this section, we will concentrate on investigating equivalences for an exploration problem. Figure 9(a) depicts a canonical marine model, with ocean, conductive sediments, thin resistive reservoir and basement. The main challenge is to recover thin reservoir and deep resistive basement simultaneously. To this end, we employed CSEM and Magnetotelluric (MT) methods. While later has large penetration depth and can recover resistive basement at depth, the former is known to be sensitive to thin resistive objects (Key 2009). To simulate a survey, single receiver was placed at the sea floor which "recorded" fields at 40 frequencies in the range of 0.001-1 Hz for MT and EM field components produced by an inline electric dipole operating at frequencies of 0.1, 1 and 5 Hz towed from receiver up to the distance of 20 km. For CSEM, we inverted amplitude and phase of $E_y, B_x$ components, whereas apparent resistivity and phase was used for MT. The data were contaminated with 3% Gaussian noise.
We used two variants of model parameterization:

(i) We assumed there are five layers (recall that the true model consists of four layers, but pretend we do not possess this information) and inverted for their conductivity and thickness. In this case, regularization matrix $L = I$.

(ii) The model was split into 65 layers with varying but fixed thickness ranging from 25 to 200 m to the depths of six kilometers. In this case, regularization matrix $L$ approximates 1st order derivative operator.

Both options have pros and cons. Parameterization (i) results in a nine-dimensional problem and thus has less uncertainty and is easier to handle numerically. Parameterization (ii) is much more flexible in terms of structures it can describe, but also significantly more ambiguous and this ambiguity needs to be mitigated by regularization. Note that log-transformed model parameters were used in inversion to preserve positivity and account for large range of admissible values. For exploration of low misfit regions, we used $n_r = 10$ restarts with random initial models and generated $n_A = 10000$ equivalent solutions using $k = 32$ (see Algorithms 2 and 3). This was done for MT, CSEM and MT+CSEM data sets independently.

The results for parameterization (i) are given in Figure 10. MT alone is completely insensitive to the thin resistive layer, but can constrain basement conductivity and with less success its depth. On the contrary, CSEM recovers thin resistor quite well, in fact this feature is present in all equivalent models, but has big uncertainty for the structure underneath. Although some sampled equivalent models do sense resistive basement, the largest majority of the equivalent models do not “see” it. Inverting both MT and CSEM data simultaneously (Figure 10c) results in better constrained model with substantially lower uncertainty throughout the model.

The results for parameterization (ii) are given in Figure 11. For this experiment, we also tried different values of $p_m$ since they incur different (and only partially overlapping) correctness sets and hence drawn models give a better idea about solution uncertainty. Similar to the previous case, only simultaneous inversion of two MT and CSEM data sets resolves both thin resistive layer and basement reliably. Interestingly, however, that among equivalent solution randomly sampled from ED, we observe few models which exhibit thin resistive layer (for MT inversion) or resistive
Figure 10. True (black dashed line), maximum posterior probability (solid black line) and equivalent models (gray lines) for MT (a), CSEM (b) and joint MT-CSEM (c) inversions for model parameterized with five layers which conductivities and thicknesses were fitted.

basement (for CSEM inversion), yet there is a big uncertainty and MAP models do not have these structures. As expected, $L_2$-norm regularization has tendency to produce artificial wiggles, particularly below and above thin resistor. These behaviour is less noticeable for $L_{1.5}$ regularization. $L_1$ does not penalize abrupt changes in conductivity, but models often look too rough and corresponding MAP solutions overestimate depth of the basement. In particular, simultaneous inversion of MT and CSEM data suggests that the depth to the basement is likely to be 6 km and generated equivalent solutions show little uncertainty to this. This is most likely related to the insufficient sensitivity of the MT data and effect of imposed $L_1$ constraints. Note that similar situation has been encountered by (Rosas-Carbajal et al. 2013) who used MCMC inference algorithm with various prior constraints.
Figure 11. True (black dashed line), maximum posteriori probability (solid black line) and equivalent models (gray lines) for MT (1st row), CSEM (2nd row) and joint MT-CSEM (3rd row) inversions for different values of $p_m$ and model parameterized with 65 layers of fixed thickness.
4.5 Comparison with MCMC algorithms

Algorithm (3) adopted to draw samples from ED, allows us to quickly build an ensemble of equivalent solutions and gives a qualitative view on solution uncertainty. However, this ensemble can be biased, especially for highly non-linear problems. To avoid that, one can use MCMC methods. The aim of this section is to show that performance of MCMC methods can be greatly improved if additional information collected during optimization stage is available.

Arguably the most popular MCMC algorithm in geophysics is random walk Metropolis-Hastings (RWMH). This algorithm allows, after some burn-in stage, to draw samples from posterior PDF (1). In addition to the curse of dimensionality (Tarantola 2005) and lack of parallelism, it is well known (Gelman et al. 2014) that if used with an uninformative prior, RWMH algorithm has difficulties in exploring multi-modal distributions. These limitations have been partially mitigated by the introduction of Delayed Rejection Adaptive Metropolis (DRAM) and parallel tempering MCMC algorithms (Haario et al. 2006; Ray et al. 2013). However, application of these methods is still challenging for high dimensional problems, and quite involved algorithms are required for further improvements (Vrugt et al. 2009). For problems with large number of unknowns where measurements are only partially informative about some model parameters, it is recommended to use informative priors (Rosas-Carbajal et al. 2013) or ad-hoc model parameterizations (Hawkins & Sambridge 2015), which can help limit class of admissible solutions.

For algorithms mentioned in this section, default recommended parameters were used without additional tuning.

4.5.1 Multivariate Gaussian distribution

To compare performance of various MCMC methods with CMAES for the problems with large number of unknowns, we took a simple problem. If algorithms show unsatisfactory behaviour on this example, their application to realistic problems will be rather limited.

First, we used RWMH and DRAM to sample from a 100-dimensional Gaussian distribution. In other words, we want to construct a set of vectors $\mathbf{x} \sim \mathcal{N}(\mathbf{\mu}, \mathbf{C})$ by using MCMC. Note, this problem is equivalent to a linear least-squares method and therefore can be solved analytically.
without resorting to MCMC, but this example serves as a good benchmark for the methods used. For this task, we chose $\mu = 0$ as a mean vector and the covariance matrix $C_G$ was constructed such that $(C_G)_{ii} = i$ and correlation between any two variables equals 0.5. We used Gaussian proposal distribution with the optimal covariance scaling $P_{prop}(\tilde{x}) = N(\tilde{x}, \frac{2\sigma^2}{M} I)$ [Gelman et al. 2014], where $M = 100$. Initial point was $x_0 = 15$. RWMH and DRAM algorithms were run for 1 000 000 iterations with the burn-in of 800 000 samples.

Further, this problem was solved by using CMAES algorithm. Upon convergence, we retrieve position of the MAP (coincides with the mean for symmetric Gaussian distribution) and covariance matrix. After matrix rescaling, one can draw samples. We used the same initial step size as for MCMC algorithms, that is $s = 2.4/\sqrt{100}$. For comparison, we also added results obtained by using Genetic Algorithm (GA) for which we employed built-in MATLAB™ 2014 implementation.

Figure 12 shows mean models calculated using last 200 000 samples of the chains produced by the RWMH and DRAM algorithms as well as MAP solutions returned by the active CMAES and GA. It is clear that 800 000 burn-in iterations were not enough for both RWMH and DRAM. None of the algorithms gave expected value within the accuracy of 10%, although chains go to the right direction. In contrast, active CMAES retrieved mean vector perfectly using a population of size 17 (equation 7). The GA was terminated after the same number of iterations as CMAES, but did not reach the MAP point. Moreover, GA required much larger population size of 200 candidates and does not provide any additional information (such as covariance matrix) rendering this method less efficient for our problems – an observation confirmed by other studies [Arsenault et al. 2013; Elshall et al. 2015].

Figure 13 shows parameter values for the 1st and 100th variables. As can be seen in Figure 13(a), RWMH produces non-stationary chains for both parameters, which means 1 000 000 iterations is not enough. This is also apparent by looking at the histograms calculated from the last 200 000 samples of the chain. The algorithm did not sample posterior PDF sufficiently well. DRAM, being more elaborated version of RWMH, produces better results. One sees in Figure 13(b) that chains reach stationary phase and oscillate around mean. However, $\approx$800 000 iterations were required to reach this stage. Corresponding normalized histograms look much better.
than for RWMH, yet some bias in the mean value is present (see also Figure 12). Finally, Figures 13(c-d) show results for the active CMAES using different population sizes. It attains good performance, practically converging to the MAP point in roughly 1000 and 400 iterations for, respectively, smaller and larger populations. The rescaled covariance matrix and MAP model returned by this method were used to draw samples from $\mathcal{N}(x_{MAP}, C_{MAP})$. It is observed that for smaller population size, CMAES produces up to 35% errors in variance estimation, which is greatly reduced after increasing population size.

It should be noted that DRAM algorithm has $O(M^3)$ time and $O(M^2)$ memory complexity, making it more expensive than classic CMAES. These limitations of MCMC methods will most likely become more pronounced and lead to severe limitations for non-linear problems with larger numbers of unknowns and expensive forward solves. In fact, sequentially performing 1 000 000 forward solves is already rather limiting requirement.
Figure 13. Variable chains and normalized histograms of 1st and 100th parameters for Random Walk Metropolis-Hastings (RWMH), Delayed Rejection Adaptive Metropolis (DRAM) and CMAES. For CMAES, mean vectors for every iteration are shown. To calculate normalized histograms, samples drawn during burn-in stage were dropped. Solid curve on histograms shows true marginal PDF. Note acceptance rate (AR) and population size ($\lambda$) values shown in the plot titles.

Note that similar test was presented by Vrugt et al. (2009). They overcome these difficulties by using very sophisticated differential evolution adaptive MCMC algorithm with multiple (200) chains.
4.5.2 Non-linear problems

To show advantages of the approach demonstrated in the previous sections for non-linear problems, we used example shown in Figure 4. A simple problem with only two unknowns exposes difficulties, which RWMH algorithm experiences when dealing with multi-modal non-linear problems.

For this problem, we first ran RWMH algorithm for 4000 iterations. Without a better prior alternative, we chose \( s = 2.4 / \sqrt{2} \) as the standard deviation for the normal proposal distribution because it is the optimal choice for Gaussian PDF (Gelman et al. 2014). Following Section 4.1, the logarithm of \( \mathbf{m}_0 = (0.1, 10) \) was used as a starting point. Figure 14(a) shows parameter chains and acceptance rate for this run. Clearly, algorithm did not perform well. Acceptance rate is too low and chains are not mixed with high auto-correlation coefficients. The vicinity of the MAP model was not sampled at all. Increasing proposal variance will make things even worse, whereas
Exploration of equivalences in inverse problems

Reducing variance will result in higher acceptance, but also higher probability of missing MAP due to non-linear topography of the posterior PDF. The only way to make RWMH work for this case is to start in the vicinity of the MAP model and choose a better proposal distribution. In real applications, this type of fine-tuning is exactly what one would like to avoid. The next paragraph shows one way to do that.

Alternatively, we can start RWMH algorithm with \( m_{\text{MAP}} \) returned by the CMAES and use \( C_{\text{MAP}} \) for proposal distribution. Figure 14(b) shows results for this hybrid algorithm, here called CMAES-RWMH. With this kind of “preconditioning”, CMAES-RWMH exhibits much better performance. Acceptance rate of 15% and good mixing indicate that posterior was sampled appropriately. Note that for realistic problems, one can start multiple MCMC chains initialized with models from \( S_{\text{low}} \), thereby allowing algorithm to sample multiple low-misfit basins found by CMAES.

5 CONCLUSIONS AND OUTLOOK

It is well known that no optimization or sampling algorithm is overall better for all types of problems (Mosegaard 2012). It is more sensible to discuss methods and approaches which may be better for a given type of problems. We have shown in this work that CMAES and modifications thereof are among such methods. It is a global optimization method, which readily allows us to use arbitrary norms for likelihood and prior terms, and requires little to no fine-tuning – default parameters worked well for majority of the studied problems. It was shown that for problems with \( L_2 \)-norm functional, the number of iterations scales at a rate of \( O(M^{1.2}) \) and the convergence can be further improved by using larger population sizes. Thanks to the embarrassingly parallel nature of the algorithm, increased runtimes due to larger population size can be compensated by using more CPUs. Furthermore, modifications of the algorithm with linear internal complexity also exist and can increase algorithm’s efficiency on larger problems.

We tested various norms used for regularization term when over-determined model parameterizations were solved. Among problems studied, \( L_1 \) norm is overall least stable and reliable. It works best on models with few distinct features, but the smooth transitions of physical properties cannot be recovered. \( L_2 \) norm produces stable results, but may over-smooth a solution and tends to cause
some artificial wiggles which may be rather misleading. $L_{1.5}$ represents an attractive trade-off and can be recommended as a default choice in practice.

Being a powerful optimization method, CMAES also provides estimates of covariance matrix for the final model. After appropriate rescaling and in combination with low misfit solutions probed during the evolution, this can be used to sample equivalence domain of the parameter space. By building an ensemble of equivalent solutions, we can estimate uncertainty and investigate non-trivial inter-dependencies in the model. Since this approach is heuristic and treats all low-misfit models equally, there is no guarantee that this ensemble is exhaustive and unbiased, i.e. uniformly samples all low misfit basins. Alternatively, one could use MCMC sampling. We have shown that efficiency of MCMC algorithms is greatly improved when these algorithms are initialized with information given by the CMAES.

Problems with large number of unknowns, which are governed by expensive to solve PDE equations pose high computational demands for stochastic methods, let alone the problem of uncertainty quantification. A combination of deterministic and stochastic approaches together with model dimension reduction techniques can make stochastic optimization and UQ feasible for these types of problems.

6 ACKNOWLEDGMENTS

Nikolaus Hansen is acknowledged for a number of insightful discussions and suggestions regarding efficient usage of CMAES. We also thank Kerry Key for making 1D CSEM modeling code Dipole1D available. Constructive comments by C. Farquharson and A. Ray helped improve the manuscript.

APPENDIX A: CMAES ALGORITHM

$(\mu/\mu_w, \lambda)$—CMAES is summarized in diagram (4). Detailed derivation can be found in Hansen & Ostermeier (2001); Hansen (2006).

The algorithm begins by initializing a number of parameters. The in-depth discussion behind
Exploration of equivalences in inverse problems

their choice is given in (Hansen & Ostermeier 2001). Note that vector \( \mathbf{m} \) contains transformed physical parameters such as log conductivity and/or log thickness.

At every iteration, the algorithm first samples from current multivariate normal distribution and evaluates minimized functional for generated realizations. These samples are then sorted according to the function values such that \( \phi(\mathbf{m}_1; \lambda) < \phi(\mathbf{m}_2; \lambda) < \cdots < \phi(\mathbf{m}_\lambda; \lambda) \). This requires solving \( \lambda \) forward problems and can be parallelized. This step is followed by selecting \( \mu \) best candidates and translating distribution mean by a displacement \( s(\langle \mathbf{m} \rangle_w) \).

Next comes step size, \( s \), adaptation. This is of critical importance in having competitive performance of the algorithm. To adapt \( s \), CMAES analyses length of the evolution path \( \mathbf{p}_s \) (that is, a sequence of successive steps taken by CMAES over a number of iterations) and tries to match it with the expected length upon random selection \( \mathbb{E}[\|\mathcal{N}(0, \mathbf{I})\|] = \sqrt{2\Gamma((M + 1)/2)/\Gamma(M/2)} \) (Hansen & Ostermeier 2001). If evolution path length is longer than expected, the step length is increased and vice versa. Note that since \( c_s < 1 \), the evolution path stores information about past steps. This information decays at the rate of \( 1/e \).

The covariance matrix update consists of two terms, namely rank-1 and rank-\( \mu \) updates. The old information in the covariance matrix fades out at the rate of \( 1 - c_1 - c_{\mu} \). The rank-one update encapsulates information about correlations between iterations, whereas rank-\( \mu \) update exploits current generation statistics.

Possible stopping criteria for Algorithm 4 are:

- Maximum allowed number of iterations, \( g_{\text{max}} \), is reached.
- Maximum allowed number of forward calculations, \( n_{f,\text{max}} \), is reached.
- Covariance matrix becomes singular, i.e. condition number \( \kappa(C_{\text{CMA}}) > 10^{14} \).
- No significant changes in the function value over last \( n_f \) iterations is observed.
- No significant changes in the mean model over last \( n_m \) iterations is observed.
Algorithm 4 \((\mu/\mu_w, \lambda)\)–CMAES algorithm

1: \(\lambda = 4 + \lceil 3 \log M \rceil, \mu = \lceil \lambda/2 \rceil\)

2: \(w_i = \frac{w_i'}{\sum_{j=1}^{\mu} w_j'}, w' = \log(\lambda/2 + 0.5) - \log i, \quad i = 1 \ldots \mu\)

3: \(\mu_{\text{eff}} = \frac{1}{\sum_{j=1}^{\mu} w_j'^2}, c_s = \frac{\mu_{\text{eff}}+2}{M+\mu_{\text{eff}}+5}, d_s = 1 + 2 \max \left(0, \sqrt{\frac{\mu_{\text{eff}}-1}{M+1}} - 1\right) + c_s\)

4: \(c_e = \frac{4+\mu_{\text{eff}}/M}{M+4+2\mu_{\text{eff}}/M}, c_1 = \frac{2}{(M+1.3)^2+\mu_{\text{eff}}}, c_\mu = \min \left(1 - c_1, \frac{2(\mu_{\text{eff}}-2+1/\mu_{\text{eff}})}{(M+2)^2+\mu_{\text{eff}}} \right)\)

5: Set evolution paths \(p_s = 0, p_c = 0\), covariance matrix \(C_{\text{CMA}} = I\)

6: Set initial mean \(\bar{m} = m_0\) and step length \(s \in \mathbb{R}_+\)

7: \textbf{for} \(g \leftarrow 1, g^{\text{max}} \) \textbf{do}

8: Draw new population:

9: \(\bar{m}_i \sim N(0, C_{\text{CMA}}), \quad i = 1, \ldots, \lambda\)

10: Evaluate \(\phi(\bar{m} + sm_i)\)

11: Select parents and update distribution mean:

12: \(\langle m \rangle_w = \sum_{i=1}^{\mu} w_i m_{i; \lambda}\)

13: \(\bar{m} \leftarrow \bar{m} + s \langle m \rangle_w\)

14: Control step size:

15: \(\bar{p}_s \leftarrow (1 - c_s)\bar{p}_s + \sqrt{c_s(2 - c_s)\mu_{\text{eff}}} C_{\text{CMA}}^{-1/2} \langle m \rangle_w\)

16: \(s \leftarrow s \times \exp \left(\frac{c_\mu}{d_s} \left(\frac{\|p_s\|}{\|p_s\|_{\text{N}(0, I)}} - 1\right)\right)\)

17: Covariance matrix adaptation:

18: \(\bar{p}_c \leftarrow (1 - c_e)\bar{p}_c + \sqrt{c_e(2 - c_e)\mu_{\text{eff}}} \langle m \rangle_w\)

19: \(C_{\text{CMA}} \leftarrow (1 - c_1 - c_\mu)C_{\text{CMA}} + c_1 \left(\begin{pmatrix} \bar{p}_s \bar{p}_c^T + \delta(h_s)C_{\text{CMA}} \end{pmatrix} + c_\mu \sum_{i=1}^{\mu} w_i m_{i; \lambda} m_{i; \lambda}^T \right)\)

REFERENCES


A. V. Grayver and A. V. Kuvshinov


Exploration of equivalences in inverse problems


Key, K., 2009. 1D inversion of multicomponent, multifrequency marine CSEM data: Methodology and synthetic studies for resolving thin resistive layers, Geophysics, 74(2), F9–F20.


A. V. Grayver and A. V. Kuvshinov


